

Descriptive Statistics Lecture

Psychology 280
Orange Coast College
2/1/2006

Definitions:

- Statistics have been defined as "a collection of methods for planning experiments, obtaining data, and then analyzing, interpreting and drawing conclusions based on the data" (Triola, 1992, p.4).
 - This definition illustrates the strong link between research methodology and statistics
 - The research design is the foundation of a good study
 - No statistic can fix an inferior research design
 - "Garbage InGarbage Out"

Definitions:

- All types of statistics can be categorized into two subgroups:
 - Descriptive statistics
 - Describe large amounts of data in an abbreviated way
 - Describe important characteristics of your data
 - Inferential Statistics
 - Goes beyond mere description
 - Use sample data to draw conclusions and make inferences about a population

Levels of Measurement

- Levels of measurement can be broken down into a hierarchy with four categories:
 - Nominal ←————— Lowest Level
 - Ordinal
 - Interval
 - Ratio ←————— Highest Level
- Statistics are appropriate or not appropriate depending on the levels of measurement of your data

Nominal Level of Measurement

- Nominal level of measurement consists of names, labels and categories.
- Subjects are classified or identified according to common characteristics.
- Nominal data are not graded, ranked, or scaled in any manner.
- Nominal measurement do not imply any quantity or direction.
- Examples: Gender, Ethnicity

Ordinal Level of Measurement

- Ordinal level of measurement goes beyond mere classification and assigns order to the values of the variable.
- Used to order or rank order objects or events
- However, the limitation of ordinal measurement is that the distances between values on the scale or continuum may not be either meaningful or known.
- In other words, ordinal scales assign order, but do not have a standardized or replicable scale
- Examples: Rankings, such as the order runners complete a race

Interval Level of Measurement

- Interval level of measurement not only classifies and give us a rank order, but also gives us information about the distances between the ranks.
- The intervals between the numbers are the same.
- However, the limitation of interval measurement is that the scale does not have a non-arbitrary or meaningful zero point. Thus, direct ratio comparisons are not possible.
- In other words, interval scales assign order, and have a standardized scale, but do not have a meaningful zero.
- Examples: Temperature in Fahrenheit or Celsius....

Ratio Level of Measurement

- Ratio level of measurement is the highest level of measurement. Therefore, these variables have all the characteristics:
 - The scale has order
 - A standard unit of measure, and
 - A meaningful zero point
- Thus, direct ratio comparisons are possible
- Examples: Salary, Number of Children, Siblings

The only difference between interval and ratio scales is the absolute zero point. They can be mathematically treated as equivalent!

Why is it important to know this?

- Statistical technique used depends on the level of data measured
- Nominal and ordinal data are discrete
 - Measured in units that cannot be divided further
 - When playing poker, how many aces can you have?
- Interval and ratio data are continuous
 - Data can be broken down into an unlimited number of values
 - The majority of traditional statistical operations require the data to be continuous

Descriptive Statistics

- Simply describing a set of data
- Can be grouped into four broad categories:
 - Tables & graphs of data
 - Measures of central tendency
 - Measures of dispersion or variance
 - Measures of position (percentile/quantile, z scores)

Descriptive Statistics

- Tables & graphs of data:
 - Many different forms of tables and graphs can be created
 - A frequency table is a list of all values found with the corresponding frequency; can display as count of numbers or percents
 - Graphs have been called the ultimate form of descriptive statisticsa picture is worth a thousand words

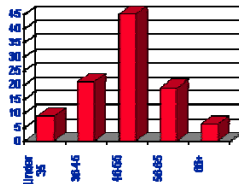
Descriptive Statistics

- Frequency distribution table of age categories

Category	Percent
Under 35	6%
36-45	21
46-55	45
56-65	19
66+	8

Descriptive Statistics

- Frequency distribution bar chart



Descriptive Statistics

- Consider an example in which we want to find out how a sample of 25 people respond to failure in an exam when asked to circle a number on the following scale:

To what extent were external factors responsible for your failure?
not at all 1 2 3 4 5 6 7 extremely

- Let's imagine their responses were as follows:
 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7

Descriptive Statistics

What do we do with this data?

- List these numbers in full every time we talk about this experiment?
- Use **statistics** to characterize the important properties of these responses — because that's what statistics are: *numerical statements about the properties of a given set of data.*
- One of the first things we might most want to say about these data is what constitutes a *typical* value.
- This is what **measures of central tendency** do.

Descriptive Statistics

- Measures of central tendency

- Describe the central scores in the data.
- These measures include:
 - Mean**
 - The mean is the arithmetic average
 - Median**
 - The median is the middle score when data are arranged in order
 - Mode**
 - The mode is the most frequently occurring score

Mean

- The mean is the *average value* (response) calculated by
 - summing all the values ($\Sigma X = 110$)
 - and dividing it by the number of values ($N = 25$).

$$\Sigma X / N = 110 / 25 = 4.40$$

Also known as the arithmetic average

$$\bar{X} = \frac{\Sigma X_i}{N} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} \quad \text{mean of a sample}$$

$$\mu = \frac{\Sigma X_i}{N} \quad \text{mean of a population set of scores}$$

Where:
 X_1, \dots, X_n = raw scores
 \bar{X} (read 'X-bar') = mean of a sample set of scores
 μ (read 'mew') = mean of a population set of scores
 Σ (read 'sigma') = summation sign
 N = number of scores

Mean (con't)

- Only appropriate when scores are on an interval or ratio scale
- Scientific abbreviation is M

Median

- The middle score when numbers are arranged in order
- If all values are ranked from 1 to N , the median is the $(N+1)/2$ value.
- If there is an ODD number, the median is the middle number
 - 2, 3, 4, 5, 5, 7, 7, 7, 8 Median = 5
- If there is an EVEN number, the median is the mean of the two middle numbers
 - 2, 3, 4, 5, 5, 7, 7, 7, 7, 8 Median = 6
- Scores must be on a ordinal, ratio or interval scale
- Scientific abbreviation is *Mdn*

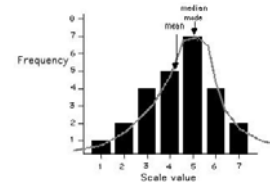
Mode

- Most frequently occurring score
 - 8, 7, 2, 5, 3, 7, 4, 5, 7
 - The mode for this distribution is 7
- A distribution can have more than one mode. This type of distribution is called multimodal.
 - 8, 7, 2, 5, 3, 7, 4, 5, 7, 5
 - The modes for this distribution are 5 and 7
- Only measure of central tendency that can be used with nominal data
- Used when only a very simple description of scores is needed (e.g., the modal age of a group of individuals).

The Relationship between Measures of Central Tendency

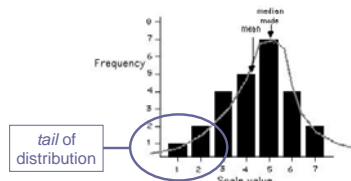
- It is clear that these three statistics — the mean, the median and the mode — *differ*.
- But do they differ in a *systematic* way?
- Yes — in fact their relationship to each other depends on the overall *distribution* of responses.
- In this example, the frequencies with which each scale point is selected can be represented in a *frequency graph* as follows....

The Relationship between Measures of Central Tendency



- The shaded bars here are based on the actual data obtained from an experimental *sample*.
- The best *estimate* of the population's behaviour (based on sample data) is represented by the curved line.

The Relationship between Measures of Central Tendency

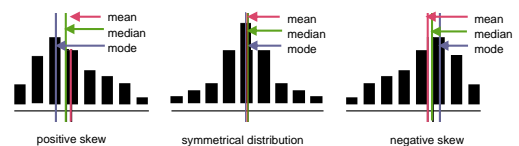


- Now in fact we can see that the distribution of responses here is slightly *skewed* — in fact it is *negatively skewed* because the distribution's *tail* extends further in the negative direction.

The Relationship between Measures of Central Tendency

When a distribution is skewed the mean, median and mode will not all be the same.

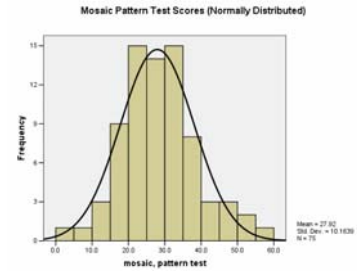
- In cases of skew, generally the *mean* is extremitized in the direction of skew more than the *median* which is extremitized in the direction of skew more than the *mode*.



The Relationship between Measures of Central Tendency

- In cases of skew, the median may be a more appropriate statistic to describe central tendency than the mean.
- This is because
 - the median falls between the mean and mode and
 - the mean can be a very misleading statistic if it differs appreciably from the median or mode.
- In fact, of the various possible distributions, one form of *symmetrical* distribution — the normal distribution (often called *the bell curve*) is the most common.
- For statistical purposes a normal distribution of responses is also the most desirable — because it has lots of useful properties (as we will see in later lectures).

Normal Distribution



The Relationship between Measures of Central Tendency

- Things like IQ and height are normally distributed.
- On the other hand, income and wealth are positively skewed (so governments often report mean income because it's higher than the median; i.e., what the average person earns).
- The ages of students in first-year university course are also positively skewed. So, if someone wanted to make students seem young, which measure of central tendency would they use?
- Self-evaluations tend to be negatively skewed. If someone wanted to emphasize the extent to which this sample was externalizing, which statistic would they use?

Comparison of Central Tendency Measures

Baseball Salaries - 2004 Anaheim Angels (Selected Players)

Player	Salary
Vladimir Guerrero	\$11,000,000
Bartolo Colon	\$11,000,000
Troy Glaus	\$9,900,000
Darin Erstad	\$7,500,000
Kelvin Escobar	\$5,750,000
Adam Kennedy	\$2,500,000
David Eckstein	\$2,150,000
Shane Halter	\$575,000
Jeff DaVanon	\$375,000
John Lackey	\$375,000
Francisco Rodriguez	\$375,000
Jose Molina	\$335,000
Chone Figgins	\$320,000
Total Salary (Sum)	\$59,155,000.00
Mean	\$4,011,928.58
Median	\$2,150,000.00
Mode	\$375,000.00

All Salaries

Source: USA Today (2004 Season)

What's Happening?

- The mean is being pulled towards the extreme high scores
- This is inflating the average and misrepresenting the data
- The median is a better choice for a measure of central tendency when scores are known to vary in extreme
- The mode may not be a useful measure for data that is continuous in nature

Measures of Dispersion

- Determine how much variability exists in a set of scores
 - Range
 - Standard Deviation
 - Variance

Range

- The range is the total number of units between the highest and lowest scores
 - Range = highest score – lowest score
- Easy to calculate but gives only a relatively crude measure of dispersion
- Only measures the spread of the two extreme scores and not the spread of the scores in between
- 8, 7, 2, 5, 3, 7, 4, 5, 7 Range = 8 – 2 = 6

Standard Deviation

- The average deviation of scores from the mean
- The standard deviation is SMALL when when the majority of scores are around the mean
- The standard deviation INCREASES as more scores lie further from the mean score
- The standard deviation (symbolized as s) is derived by first calculating the variance (symbolized as s^2)
 - It is the square root of the variance
- Similar to the mean, only appropriate for interval and ratio level scores

Standard Deviation

- Gives us a measure of dispersion relative to the mean
- Is sensitive to each score in the distribution
- Like the mean, the standard deviation is stable with regard to sampling fluctuations
- Both the mean and standard deviation can be manipulated algebraically and allows them to be used in inferential statistics calculations

Standard Deviation

- Standard deviation tells us a lot about a distribution, particularly if that distribution is normally distributed.
- If this is the case, we know for example, that about 68% of all values will fall within 1 SD of the mean, 95% fall within 2 SDs and 99% fall within 3 SDs.
- As we will see in later lectures, this type of information is very useful to know.

The Principles of Variance

- The basic concept is to measure how scores vary about the mean
- So logic says, take each score and subtract from the mean, then sum the difference and make an average
- But what happens? The sum will always be ZERO!
- To fix this, square the difference, then average. Thus, you've arrived at the variance and it's a squared unit of measurement
- The main problem with variance as a measure of dispersion is that it is not given in units of X but in units of X^2 .
 - So if we were measuring height, variance would be in units of height² (i.e., area)
- Our fix is the standard deviationtake the square root of the variance.

Variance & Standard Deviation

Here's the statistical formula for variance:

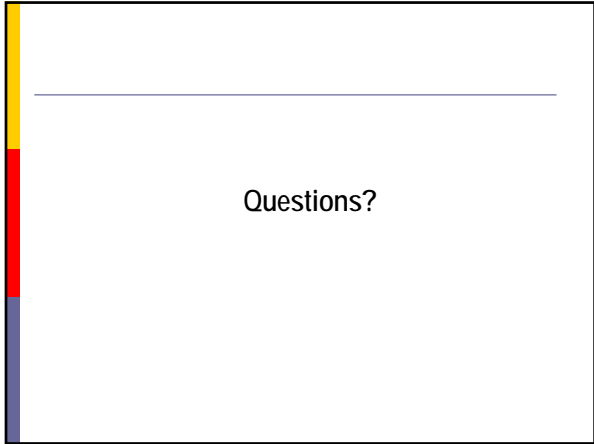
$$S_x^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$$

Here's the statistical formula for standard deviation:

$$S_x = \sqrt{S_x^2}$$

You might ask why divide by N rather than $N-1$?

- Sample data tend to *under-estimate* population variance, we need to make a slight adjustment to this statistic.
- We do this by dividing the numerator of the equations for variance and standard deviation by $N-1$ rather than N .



Questions?